MAGE: MAsked Generative Encoder to Unify Representation Learning and Image Synthesis

Year: 2022 | Citations: 220 | Authors: Tianhong Li, Huiwen Chang, Shlok Kumar Mishra, Han Zhang, D. Katabi

---

Abstract

Generative modeling and representation learning are two key tasks in computer vision. However, these models are typically trained independently, which ignores the potential for each task to help the other, and leads to training and model maintenance overheads. In this work, we propose MAsked Generative Encoder (MAGE), the first framework to unify SOTA image generation and self-supervised representation learning. Our key insight is that using variable masking ratios in masked image modeling pre-training can allow generative training (very high masking ratio) and representation learning (lower masking ratio) under the same training framework. Inspired by previous generative models, MAGE uses semantic tokens learned by a vector-quantized GAN at inputs and outputs, combining this with masking. We can further improve the representation by adding a contrastive loss to the encoder output. We extensively evaluate the generation and representation learning capabilities of MAGE. On ImageNet-1K, a single MAGE ViT-L model obtains 9.10 FID in the task of class-unconditional image generation and 78.9% top-1 accuracy for linear probing, achieving state-of-the-art performance in both image generation and representation learning. Code is available at https://github.com/LTH14/mage.